Defense Announcement

Unsupervised Discovery and Representation of Subspace Trends in Massive Biomedical Dataset

Yan Xu

Degree: PhD, Electrical Engineering

Date: 12/02/2015 **Time:** 1:00PM

Location: N325 Engineering Building 1

Committee Chair: Dr. Badri Roysam, ECE

Committee Members: Dr. David Mayerich, ECE
Dr. Jared Burks, MD Anderson Cancer Center
Dr. Peng Qiu, Georgia Tech
Dr. Saurabh Prasad, ECE
Dr. Zhu Han, ECE

This goal of this thesis is to develop unsupervised algorithms for discovering previously unknown subspace trends in massive multivariate biomedical data sets without the benefit of prior information. A subspace trend is a sustained pattern of gradual/progressive changes within an unknown subset of feature dimensions. A fundamental challenge to subspace trend discovery is the presence of irrelevant data dimensions, noise, outliers, and confusion from multiple subspace trends driven by independent factors that are mixed in with each other. These factors can obscure the trends in traditional dimension reduction and projection based data visualizations. To overcome these limitations, we propose a novel graph-theoretic neighborhood similarity measure for sensing concordant progressive changes across data dimensions. Using this measure, we present an unsupervised algorithm for trend-relevant feature selection and visualization. Additionally, we propose to use an efficient online density-based representation to make the algorithm scalable for massive datasets.

The representation not only assists in trend discovery, but also in cluster detection including rare populations. Our method has been successfully applied to diverse synthetic and real-world biomedical datasets, such as gene expression microarray and arbor morphology of neurons and microglia in brain tissue. Derived representations revealed biologically meaningful hidden subspace trend(s) that were obscured by irrelevant features and noise. Although our applications are mostly from the biomedical domain, the proposed algorithm is broadly applicable to exploratory analysis of high-dimensional data including visualization, hypothesis generation, knowledge discovery, and prediction in diverse other applications.